Étiquetage morpho-syntaxique pour des mots nouveaux

Ingrid Falk & Delphine Bernhard & Christophe Gérard & Romain Potier-Ferry

lilpa linguistique, langues, parole

UNIVERSITÉ DE STRASBOURG

Problématique : quels étiqueteurs pour les néologismes formels ?

Projet : **Logoscope** (2012/15)

- détection automatique de néologismes (presse fr.)
- documentation néographique

Besoin : un outil. . .

- performant pour l'étiquetage de néologismes
- prêt à l'emploi : sans entraînement supplémentaire
- simple d'utilisation
- librement disponible

Objectifs de l'étude :

- comparaison de 7 étiqueteurs pour le français
- étiquetage de néologismes issus du Wiktionnaire
- analyse de leurs performances

Corpus de référence : néologismes extraits du Wiktionnaire

- néologismes attestés (pas forgés)
- contextualisés (phrase ou paragraphe)
- apparition postérieure à 2000

Format

```
<entry id="L00003"><word pos="verbe">accessibiliser</word>
  <examples>
    <texte url="http://...." date="2003" type="blog" id="L00003-1">
<neologisme>Accessibiliser</neologisme> l'Internet... oui, mais pas à moitié!</texte>
    <texte url="http://..." date="2009" type="blog" id="L00003-2">
Ne pas passer une journée à tenter d'<neologisme>accessibiliser</neologisme> deux-trois tableaux
hallucinants, pour un gain trop faible.</texte>
  </examples>
</entry>
```

Informations utilisées

- lemme (`<word>`)
- catégorie grammaticale (attribut `pos`)
- occurrence (`<texte>`)

Résumé

#	total	noms	verbes	adjectifs	adverbes	locutions	mots à tiret
lemmes	158	84	36	29	3	6	11
occurrences	459	293	68	81	4	13	28
phrases	220	115	55	38	4	8	19

Étiqueteurs morpho-syntaxiques utilisés

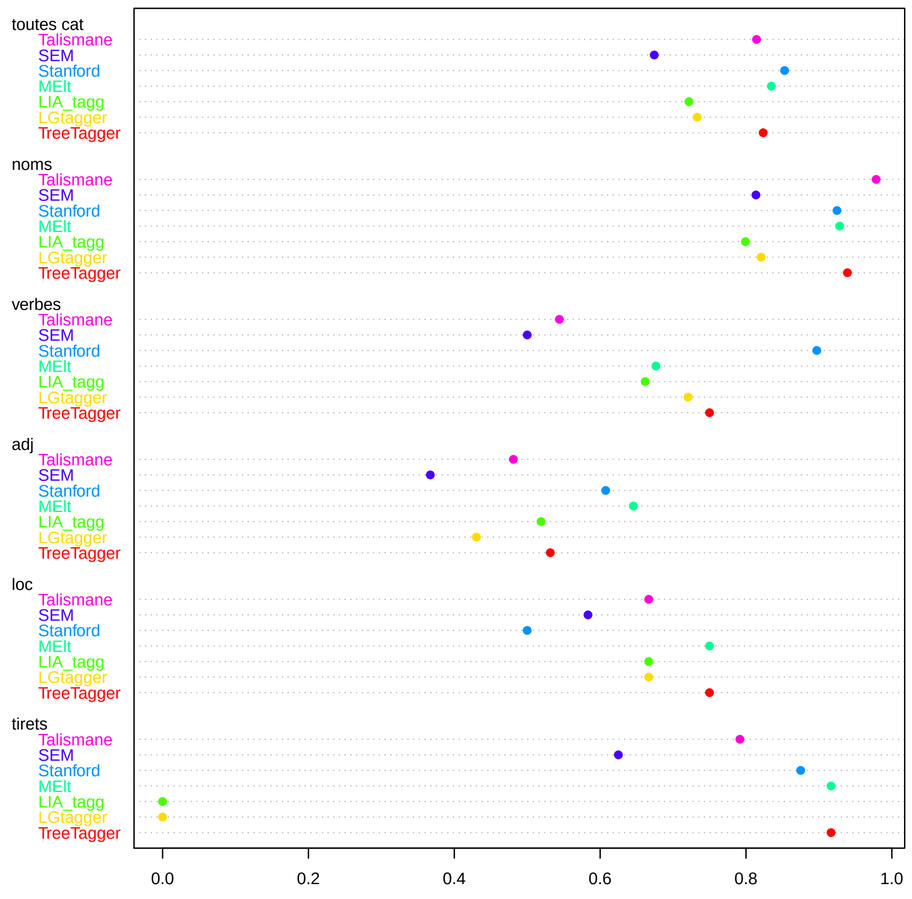
Outil	Méthode	Corpus d'apprentissage (étiquettes)	Ressource lexicale	Utilisation de la forme	Particularité
LGTagger	CRF	FTB (?)	DELA, Lefff, Prolex, Organisations, Prénoms	Oui (i)	segmentation incorporée
SEM	CRF	FTB (?)	Lefff	Oui (i)	
LIA_tagg	HMM	? (103)	lexique 10 000 mots	Non	
Stanford	CMM à maximisation d'entropie.	FTB (14)	–	Oui (i)	bidirectionnel
MElt	CMM à maximisation d'entropie.	FTB (29)	Lefff	Oui (i)	
Talismane	EM	FTB (?)	Lefff	Oui	
TreeTagger	arbres décision	? 43 834 mots (33)	–	Oui (ii)	

(i) Utilisation de traits « mots inconnus » : *n*-gram suffixes et préfixes (*n* de 1 à 4), tirets, majuscules, chiffres, etc.

(ii) Utilise un lexique associant à des suffixes la probabilité des étiquettes.

Résultats : % étiquettes correctes

Étiqueteur	toutes	noms (293)	verbes (68)	adjectifs (81)	locutions (13)	mots à tiret (28)
LGtagger	73.30	82.08	72.06	43.04	66.67	0.00
LIA_tagg	72.17	79.93	66.18	51.90	66.67	0.00
MElt	83.26	92.83	67.65	**64.56**	**75.00**	**91.67**
SEM	67.42	81.36	50.00	36.71	58.33	62.50
Stanford	**85.29**	92.47	**89.71**	60.76	50.00	87.50
Talismane	81.45	**97.85**	54.41	48.10	66.67	79.17
TreeTagger	82.35	93.91	75.00	53.16	**75.00**	**91.67**
majorité	86.43					



Bilan

- meilleures performances : Stanford > MElt > TreeTagger
- propriétés favorables : traits de forme et morphologiques
- propriétés peu utiles : ressource lexicale, segmentation-étiquetage simultané

Travaux financés par l'Université de Strasbourg dans le cadre de l'Initiative d'Excellence (IdEx) 2012-2015 (projet **Logoscope**)

{ifalk,dbernhard,christophegerard}@unistra.fr ; romainpotierferry@gmail.com